DOCUMENT RESUME

ED 235 235 TM 830 662

AUTHOR Bowman, Robert W., Jr.; Frary, Robert B.
TITLE Difficulty Level of Classroom Tests: Communicating with Teachers and Students.
PUB DATE Apr 83
NOTE 45p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education, (Montreal, Quebec, April 12-14, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; *Difficulty Level; Higher Education; Measurement Techniques; Norm Referenced Tests; Questionnaires; *Student Evaluation; *Teacher Made Tests; *Test Construction; Test Format; *Testing Problems; Test Interpretation

ABSTRACT

College teachers often use norm-referenced classroom tests which are too easy for distinguishing adequately among levels of student achievement, yet they are reluctant to adopt more difficult tests. We explored the basis for current practices concerning test difficulty through informal interviews and questionnaires completed by faculty members and by students. Overall, the responses suggest that many faculty and students view the percentage-correct score as an absolute measure, even when tests are clearly norm-referenced. Further, faculty often claimed that students would not tolerate more difficult tests even if grade distributions remained the same. Finally, we found specific support for more difficult tests in only a few of a selection of 19 measurement textbooks we reviewed. In this context we developed a booklet ["College Achievement Testing: A Self-Instructional Interlude," included as appendix A] and a pamphlet ["The Trouble with Headaches," included as appendix B] intended to convince students that norm-referenced tests should be reasonably difficult. Evaluation of these materials, however, showed them to be largely ineffective for this purpose. Apparently, beliefs in this area are not easily changed, which warrants ongoing efforts to inform teachers and students of appropriate practice regarding test difficulty. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED235235

Difficulty Level of Classroom Tests: Communicating with Teachers and Students

Robert W. Bowman, Jr., and Robert B. Frary

Virginia Polytechnic Institute & State University

Blacksburg, Virginia 24061

Paper presented at the Annual Meeting of

the National Council on Measurement in Education,

Montreal, Quebec, Canada,

April 11-15, 1983

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. W. Bowman, Jr.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 830 662

Abstract

College teachers often use norm-referenced classroom tests which are too easy for distinguishing adequately among levels of student achievement, yet they are reluctant to adopt more difficult tests. We explored the basis for current practices concerning test difficulty through informal interviews and questionnaires completed by faculty members and by students. Overall, the responses suggest that many faculty and students view the percentage-correct score as an absolute measure, even when tests are clearly norm-referenced. Further, faculty often claimed that students would not tolerate more difficult tests even if grade distributions remained the same. Finally, we found specific support for more difficult tests in only a few of a selection of 19 measurement textbooks we reviewed. In this context we developed a booklet and a pamphlet intended to convince students that norm-referenced tests should be reasonably difficult. Evaluation of these materials, however, showed them to be largely ineffective for this purpose. Apparently, beliefs in this area are not easily changed, which warrants ongoing efforts to inform teachers and students of appropriate practice regarding test difficulty.

Difficulty Level of Classroom Tests: Communicating with Teachers and Students

Robert W. Bowman, Jr., and Robert B. Frary[1]

Virginia Polytechnic Institute and State University

This paper concerns a widespread but apparently little-noticed deficiency in norm-referenced achievement testing by classroom teachers, namely, use of tests with inappropriate difficulty levels. Our work within the Learning Resources Center, a comprehensive instructional support service, involves consulting with faculty members about evaluation and testing. In our work--which includes giving seminars on testing practices, conferring about specific testing problems, and providing test scoring and analysis services--we became aware that, in a high proportion of tests, over 90 per cent of the students were answering more than 70 per cent of the items correctly.

---

[1]We wish to acknowledge with gratitude the assistance in this study of Drs. Lois Gurel, Wayne Hensley and Gordon Sabine. Correspondence concerning this paper, including requests for copies of the instructional presentations which are described, should be sent to the authors, Learning Resources Center, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

Moreover, with rare exceptions, these tests could not in any way be described as criterion-referenced. They were simply collections of questions that more or less covered the subject matter taught. Typically this coverage was very broad, consisting of small numbers of questions for each of a wide variety of topics. Clearly, resulting scores provided no more than ranking information.

Within this environment, large proportions of items were being answered correctly by more than 90 per cent of the students. Clearly, this preponderance of easy items contributes very little towards faculty efforts to distinguish levels of student achievement for the awarding of grades, a major purpose, if not the sole purpose, for testing in many courses. Improved levels of test difficulty should provide substantial educational benefit. Simple elimination of easy items alone should yield meaningful savings of paper, secretarial time and faculty time with little or no loss in testing effectiveness. More difficult tests of the original length could yield wider score distributions with concurrent improvement of ranking accuracy. Also, use of more difficult questions should increase the potential for feedback to instructors concerning student achievement in critical course areas.

This paper documents the basis for current practices with respect to test difficulty and reports our efforts to convince faculty members and students of the desirability of more difficult norm-referenced classroom tests. This effort involved

two instructional presentations, the development and evaluation of which we also describe.

The Basis for Current Practices

In 1978, the Center published a series of Testing Memos to provide faculty members with information about various aspects of testing. One of these was devoted to test difficulty, and its circulation gave rise to many discussions with faculty members who used the Center's test scoring and analysis service. Many expressed surprise and even dismay at the suggestion that a 70% passing level was much too high for most college tests. In almost all of these cases, the faculty members viewed their tests as essentially criterion-referenced. Careful elaboration of the content of the Testing Memo usually convinced them that their tests yielded only ranking information, but most said that they would continue to give relatively easy tests.

Their reasons were varied: Tradition demanded it; their tests should be similar to those they experienced as students. The students would be upset or discouraged and might deliver lower-than-usual faculty ratings. Their colleagues or department heads might accuse them of lowering standards. Some argued that the widespread use of easy tests somehow proved that this practice must be optimal.

Distribution of the Testing Memos, including the one on test difficulty has continued at Virginia Tech, and, since 1978, we have presented a number of seminars to inform faculty members

concerning desirable testing practices. With respect to test difficulty, these efforts often yielded reactions similar to those described above. Nevertheless, it seemed possible that after four years of efforts, difficult tests might enjoy some prevalence. Accordingly, data were collected on 136 multiple-choice tests brought in for our scoring and analysis service during a period of nine working days during the fall quarter, 1982. These results are summarized as follows:

Number of tests	136
Mean of mean % correct	73.2%
Mean of standard deviations of % correct	11.5%
Maximum mean % correct	93.7%
Minimum mean % correct	41.0%
Mean no. of examinees	91.6
Mean no. of items	32.6

Thus a typical test had a mean of 73% and a standard deviation of 12%, hardly very encouraging considering our efforts. Thirty-four of the 137 tests had means above 80%, and only 11 had means below 60%.

A second approach to characterizing current practices was through use of a questionnaire. Having noted prevalent comments about test difficulty and related matters over many interviews and conversations, we were able to generate easily a number of germane statements for response on a four-point scale ranging from agree to disagree. After editorial assistance from colleagues, we mailed these to two random samples of 100 faculty

members each. One sample was of faculty members who use the test scoring and analysis service or attended the seminars described above. The members of this population (N=500) had at least received the Testing Memo on test difficulty and many had received extensive encouragement to use more difficult tests. The other sample represented the remainder of faculty members (N=1000) who we believe were much less likely to have been sensitized with respect to the test difficulty issue.

The overall return rate was 54%, somewhat above average for a mailed questionnaire of this type on our campus, and each sample had a return rate over 50%. Just as there was little difference in return rate, so there was little difference between samples in response level for any item of the questionnaire. In fact, the correlation between mean response levels for the two groups over the 16 items was .96, and the average difference in response level on an item was .015 on a scale ranging from 1 to 4. Accordingly all comments which follow are based on the total return of 107 questionnaires.

Table 1 lists the questionnaire items and the percentages agreeing or tending to agree with each item. Some interesting apparent inconsistencies may be observed. While 97% agree that there should be items to challenge better students (statement 9), 51% agree that it is unfair to ask questions some students can't answer regardless of how hard they studied (statement 1). Responses to several statements suggest that a substantial proportion of faculty members view the percentage correct as some

sort of absolute or criterion-referenced measure, e.g., statements 4, 5, 6, 7, 15, and especially 11. Yet 89% agree that the difficulty of a test question depends as much on how it is asked as on its actual content (statement 13).

Insert Table 1 about here.

While there may be interpretations of the statements that might justify some of the apparent inconsistencies, the response level for one statement stands alone as an indicator of the central problem of this paper. That statement is number 5, "A student who gets 70% or more right on a test should get a passing grade regardless of the class average." Seventy-six percent of responders agreed or tended to agree with this statement.

To investigate student beliefs related to test difficulty, a second questionnaire was developed. It contained seven of the statements used on the faculty questionnaire, eight statements about test difficulty more pertinent to students and ten questions probing for reasons that harder tests are desirable for typical college classroom testing. It was administered initially in a 99-student section of a sophomore-level elective general education course enrolling students from many majors and colleges within the University. Table 2a contains the 15 opinion items of this questionnaire and the percentage agreeing or tending to agree with each statement. For the seven statements identical to those on the faculty questionnaire, the percentages of faculty agreement are shown in parentheses. Table 2b lists the latter 10 questions and gives the percentages answering each "correctly."

Insert Tables 2a and 2b about here.

On six of the seven statements common to both questionnaires, the student responses were substantially more often than those of faculty members in the direction of viewing scores from a typical college test as absolute or criterion-referenced. Only on statement 12 concerning interpretation of a score of 90% did students appear more realistic than faculty members. Statement 3, a general condemnation of tests was included to check the extent to which student responses might reflect antipathy toward testing. The proportion agreeing with this statement was only 28%, and, if the student agreement rates on statements 1, 5, 7, 11, 13, and 15 are reduced by this amount, they become very close to the faculty agreement rates on the same items.

On the latter 10 questions the students displayed apparent inconsistencies in a manner similar to that of the faculty responders. For example, even though 61% recognized that typical college test scores rank examinees rather than reveal a percentage of material known (Question 19), 94% believed that easy material should be covered in tests "because it makes up a substantial percentage of what is to be learned" (Question 18).

Both the faculty and student responses were analyzed further to determine their psychometric adequacy. Responses to the agree-disagree statements were valued integrally (1-4) and summed

(reversing scaling for statement 9 for the faculty and statement 2 for the students). Total scores thus determined had moderate degrees of internal consistency; coefficient alpha was .566 for faculty scores and .521 for student scores. Only responses to statements 9 and 13 on the faculty questionnaire failed to correlate significantly ($p < .05$) with total scores. The latter 10 questions of the student questionnaire yielded a KR20 of only .31 when scored right-wrong. However, responses to all of these questions correlated significantly with total scores. The correlation between total scores on the first 15 and latter 10 questions of the student questionnaire was -.29, indicating a weak but significant ($p < .05$) tendency for those with opinions leaning toward the absolute, criterion-referenced interpretation of test scores to answer fewer of the latter, more factual questions "correctly."

The next step in this research was to review a selection of measurement textbooks oriented towards classroom teachers. The selection was not intended to be up to date or exhaustive, but to represent texts which might have influenced practice with respect to test difficulty over the years. Even though few college instructors read these texts, we believe they may influence testing at lower levels and thus establish practices that carry over into college testing. We selected 19 textbooks for review. In each text, we located the authors' treatment of test or item difficulty, and we then analyzed each treatment, following as a general guide, Davies' (1971) procedures for topic analysis. The outcome of this procedure can be seen in Table 3.

Insert Table 3 about here.

All of the textbooks mentioned test or item difficulty as an important consideration in test construction. Further, most of the textbooks discussed the decision the test maker wants to make as influencing desired item and test difficulty, the effect of student ability on difficulty, the function of items of intermediate difficulty in producing a wide distribution of scores to maximize discrimination among students, and disadvantages of items which are too easy or too hard.

The most commonly encountered subtopic was the recommendation that test items should be of moderate difficulty. In some form, this recommendation was made in 16 of the 19 textbooks reviewed. While the recommendation varied in detail--some said an item difficulty level of 50 per cent should be sought, and others modified the recommendation in the light of guessing and the effects of item intercorrelations--the textbooks generally agreed that average item difficulties should fall at about 50 per cent.

Despite the frequency of this recommendation, only eight textbooks spelled out its consequence, namely, that the mean of the test will also fall at around 50 per cent correct (Ebel, 1972; Gronlund, 1976, 1977; Marshall & Hales, 1971; Noll, 1965; Sax, 1974; Stanley & Hopkins, 1972; and TenBrink, 1974). Two textbooks went on to anticipate common objections to such a result--that tests with means so low reflect poor achievement on

the part of students or ineffective teaching--and argued against the prevailing notion that a score of 70 per cent should mark the minimally passing score (Ebel, 1972; and Stanley & Hopkins, 1972). Only one text made the explicit point that, for a marginal student to attain 70 per cent, at least half the questions must be very easy, and that this is wasteful if not counterproductive in the case of norm-referenced classroom tests (Ebel, 1972).

Given this state of affairs--widespread lack of specific support from measurement texts for more difficult classroom tests and claims by teachers that students would neither tolerate nor understand the use of more difficult tests even if grade distributions remained the same--we chose to aim our change efforts at the students. We reasoned that if we could convince students of the need for more difficult tests then their teachers might be persuaded more easily.

Attempts to Convince Students that Tests Should Be More Difficult

We performed a task analysis (Gagné & Briggs, 1974) to assist us in selecting and sequencing learning activities important to the test difficulty issue. Based on the analysis and upon what we had learned about the views of faculty and students, we developed a printed self-instructional presentation. Taking the form of a booklet, the presentation was designed to capture student interest and to enable the student to distinguish

between criterion- and norm-referenced tests, correctly interpret percentage-correct scores in both criterion- and norm-referenced situations, identify factors which can determine test item difficulty in each situation, recognize the relationship between test difficulty and discrimination, and recognize the effects of appropriate and inappropriate levels of test difficulty on grading decisions. The resulting booklet contained a brief allegory as an introduction followed by a series of questions, answers, and explanations in a tongue-in-cheek programmed instruction format.

We then solicited reactions to the booklet from students in an undergraduate analytical writing course. As a class exercise, the teacher in the course asked the students to read the booklet and to write critiques of it. We also asked a small group of university teachers to review the booklet. On the basis of the students' and teachers' criticisms we shortened the booklet, improved some of the mechanics of the programming, and labelled the figures more clearly.

Students in a sophomore-level elective general education course (N=66) were asked to read the booklet as part of their homework and then to respond, in class, to the 25-item questionnaire described above. Our plan was to compare the responses of the students who had read the booklet with those of the students who had not. By evaluating scale and test score differences at both the total score and item level we hoped to make inferences about the effect of reading the booklet.

However, we learned from the teacher of the students assigned to read the booklet that only ten had actually done so. Apparently some of the students found the booklet to be unattractively packaged, and others simply found it to be too long. (We did compare the responses of the ten students who said they had read the booklet with the responses of the 56 who had not, and there were no systematic differences.)

Therefore, we decided to revise the booklet extensively. We asked a communications consultant for assistance in making the revision. She suggested that the presentation run no more than one page, start from the student's point of view, and employ humor, color and illustrations. The revised presentation was a pamphlet called "The Trouble with Headaches."

Students in another sophomore-level general education course (N=84) were asked to read the pamphlet in class and then to respond to the questionnaire. While this time it appeared that the students read the presentation, comparison of the responses of these students with the responses of the students who had not (N=99) revealed no differences of substance: The mean response level across the 15 agree-disagree statements (scored 1-4) was 2.20 for the treated group and 2.18 for the untreated group. The mean number right for the latter ten questions was 4.6 for the treated group and 4.2 for the untreated. We therefore concluded that our efforts to convince students that tests should be more difficult were largely unsuccessful.

At various points in the activities described above, we talked with students individually and in groups about the question of test difficulty. In these discussions we sought clues as to what beliefs or opinions might underlie the tenacity of opinions in this area. Many reactions were essentially irrelevant or illogical. However, we soon recognized two themes on which many student comments were based.

The first was essentially technical and concerned the premise that the same proportion of each letter grade would be given regardless of the score distribution. Some students expressed distrust of faculty members in this regard believing that many would use the opportunity to give progressively lower grades. A sort of mirror-image or corollary of this complaint was the observation that over time students would probably fare better under conventional testing/grading procedures: Sometimes an instructor might make the test too easy but nevertheless be required to award an excess of high grades based on preestablished standards. At other times, when a test might be too hard, the instructor would recognize the problem and "curve" the grades.

The second theme was more personal and seemed to reflect a longing for criterion-referenced measures. One student said, "I know the tests really just produce ranking information, but I'm trying to learn the material, not compete with the rest of the class and I want the test to tell me how well I did." A variation on this theme was dismay due to a perceived

inconsistency between satisfactory learning and inability to answer a large proportion of test questions. In this regard, another student said, "It's just too upsetting to find question after question I can't answer when I know I learned a large percentage of what was taught."

Discussion

It would appear that there is no simple "fix" for the problem we have investigated. Apparently testing procedures are like deeply ingrained folkways that change only slowly. Nevertheless, change is possible. We gain some inspiration from a statement made by Horace Mann in the mid-nineteenth century:

> We venture to predict that the mode of examination, by printed questions and written answers, will constitute a new era in the history of our schools.

At the time of Mann's statement, nearly all testing was oral, and students were completely at the mercy of their examiners. Only essays and reports constituted verifiable evidence of student achievement and then, as now, these were subject to severe lack of objectivity in grading.

If we have been able to introduce objectivity into educational measurement to the extent that we have, further progress is certainly possible. Locally, we know that we have large numbers of instructors who now feel less guilty about "curving" their grades. Some regularly adopt devices to circumvent the problem of too many easy questions. One instructor adds 20% to every score on every test. His scores

resemble a conventional distribution with a 70% passing point, and the students seem satisfied. Another gives a small number of very easy questions that count quintuple. He justifies this by observing that these cover the most important and basic concepts. A small number of instructors view their test scores as only rankings and use appropriate difficulty levels. One such instructor has told us that she found "The Trouble with Headaches" quite helpful in explaining her testing philosophy to her classes.

In conclusion, we recommend that measurement specialists in schools and colleges establish communication efforts to inform instructors of appropriate practice with respect to test difficulty. This effort should complement other informational presentations on improvement of measurement. We hope, also, that measurement texts for teachers will treat the test difficulty issue more thoroughly in future editions. For the present we will grant, on request, reprint rights for "The Trouble with Headaches" free of charge to those who believe this pamphlet may be useful. The "self-instructional" booklet is not copyrighted and may be reproduced if credit for authorship is given.

References

Ahmann, J. S., and Glock, M. D. Evaluating Pupil Growth: Principles of Tests and Measurements (Fourth Edition). Boston: Allyn and Bacon, 1971.

Anastasi, A. Psychological Testing (Third Edition). New York: MacMillan, 1968.

Chase, C. I. Measurement for Educational Evaluation. Reading, MA: Addison-Wesley, 1974.

Cronbach, L. J. Essentials of Psychological Testing (Second Edition). New York: Harper & Row, 1960.

Davies, I. K. The Management of Learning. London: McGraw-Hill, 1971.

Ebel, R. L. Essentials of Educational Measurement. Englewood Cliffs, NJ: Prentice-Hall, 1972.

Gagné, R. M., & Briggs, L. J. Principles of Instructional Design. New York: Holt, Rinehart & Winston, 1974.

Gerberich, J. R., Greene, H. A., & Jorgenson, A. N. Measurement and Evaluation in the Modern School. New York: McKay, 1962.

Gronlund, N. E. Measurement and Evaluation in Teaching (Third Edition). New York: MacMillan, 1976.

Gronlund, N. E. Constructing Achievement Tests (Second Edition). Englewood Cliffs, NJ: Prenctice-Hall, 1977.

Horrocks, J. E., & Schoonover, T. I. Measurement for Teachers. Columbus, OH: Charles E. Merrill, 1968.

Lien, A. J. Measurement and Evaluation of Learning (Third Edition). Dubuque, IA: William C. Brown, 1976.

Marshall, J. C., & Hales, L. W. Classroom Test Construction. Reading, MA: Addison-Wesley, 1971.

Martuza, V. R. Applying Norm-Referenced and Criterion-Referenced Measurement in Instruction. Boston: Allyn and Bacon, 1977.

Mehrens, W. A., & Lehmann, I. J. Measurement and Evaluation in Education and Psychology. New York: Holt, Rinehart and Winston, 1973.

Noll, V. H. Educational Measurement (Second Edition). Boston: Houghton Mifflin, 1965.

Sax, G. Principles of Educational Measurement and Evaluation. Belmont, CA: Wadsworth, 1974.

Stanley, J. C., & Hopkins, K. D. Educational and Psychological Measurement and Evaluation. Englewood Cliffs, NJ: Prentice-Hall, 1972.

TenBrink, T. D. Evaluation: A Practical Guide for Teachers. New York: McGraw-Hill, 1974.

Thorndike, R. L., & Hagen, E. Measurement and Evaluation in Psychology and Education (Second Edition). New York: John Wiley & Sons, 1961.

Tuckman, B. W. Measuring Educational Outcomes. New York: Harcourt Brace Jovanovich, 1975.

Table 1 - Faculty Questionnaire Results

N = 107

Response Instructions

Please answer the following questions with respect to a typical college test *other than an essay test.* For example, such a test might be multiple-choice, short answer, or involve quantitative problem solving.

Response Categories

1. Agree 2. Tend to agree 3. Tend to Disagree 4. Disagree

Statements	Percentage Agreeing or Tending to Agree
1. It's basically unfair to ask exam questions that some students can't answer no matter how hard they've studied.	51
2. Using a test to rank student achievement assures that one person's success is another person's failure.	18
3. Instructors who give difficult tests are likely to experience student hostility, regardless of how grades are assigned.	55
4. A truly good test tells the extent to which students have mastered what they have been taught.	96
5. A student who gets 70% of more right on a test should get a passing grade regardless of the class average.	76
6. Passing a student who answered correctly only 50% of the questions on a test is a violation of any reasonable academic standards.	30
7. Tests should be designed so that most marginal students who have put forth good effort can get 70% right.	34
8. Ranking students on the basis of test scores to assign grades leads to competition alien to the real purposes of education.	29
9. A good test should include some difficult questions to challenge the better students.	97
10. A test with an average score of 50% is undesirable no matter how grades are awarded.	51
11. A score of 90% usually means that the student has learned about 90% of the material covered by the test.	42
12. A good test will reveal each student's personal growth and development during a course of study.	62
13. The difficulty of a test question depends as much on how it is asked as on its actual content.	89
14. Difficult tests only discourage students, even when relatively low scores translate into acceptable grades.	51
15. When a student's score is lowest in the class, the corresponding grade should depend mainly on the percentage answered correctly.	54
16. On a truly fair test, nearly every student can succeed.	40

Table 2A - Student Questionnaire Results
N = 99

Response Instructions: Statements 1-15

Listed below are a number of statements about tests and examinations. Please read each statement carefully and answer by marking one of the following choices in the answer column:

1. Agree 2. Tend to agree 3. Tend to Disagree 4. Disagree

Statements 1-15	Percentage Agreeing or Tending to Agree [1]	
1. It's basically unfair to have exam questions that some students can't answer no matter how hard they've studied.	84	(51)
2. Multiple-choice questions, to be effective, need answer choices which will seem highly confusing to someone who did not study for the test.	48	
3. Exams are inherently unfair; grades should be based only on performance in real-life situations, such as preparing reports or solving problems.	28	
4. The only really fair exams are those for which all the answers can be found in the text, lecture notes or other course materials.	77	
5. Anyone who gets 70% or more right on a test should get a passing grade regardless of the class average.	96	(76)
6. Exams on which partial credit is given are inherently more fair than multiple-choice exams.	87	
7. A test with an average score of 50% is undesirable no matter how grades are awarded.	67	(51)
8. Open book exams are unfair when they have questions that can't be answered directly from the book.	36	
9. Take-home exams are highly desirable and should be used more widely.	56	
10. Multiple-choice and true-false exams are undesirable because someone who half knows the material and is lucky can get a good grade.	32	
11. Difficult tests only discourage students, even when relatively low scores translate into acceptable grades.	88	(51)
12. A score of 90% usually means that the student has learned about 90% of the material covered by the test.	28	(42)
13. Using a test to rank student achievement assures that one person's success is another person's failure.	48	(18)
14. It is not fair when all the choices of a multiple choice question have elements of correctness and the student has to pick the best answer.	90	
15. On a truly fair test, nearly every student can succeed.	51	(40)

[1] Figures in parentheses are percentages of faculty responders agreeing or tending to agree to same statements (see Table 1).

Table 2B - Student Questionnaire Results

N=99

Response Instructions: Questions 16-25

Following are some questions about tests and examinations. Read each question carefully, and then mark the one best answer in the answer column.

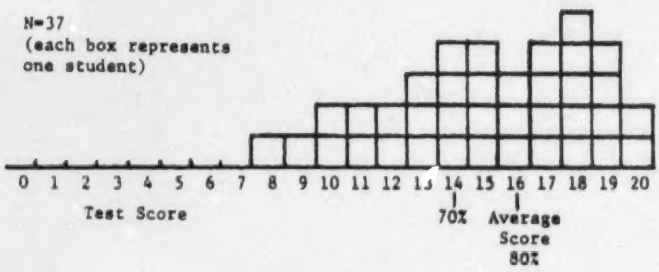
Questions 16-25	Percentage "Correct"[1]
16. Professors X and Y teach sections of the same course at Virginia Tech and cover approximately the same material in their classes. They produce their final exams independently. The average score in Professor X's class is 80% and in Professor Y's class 70%. What is the most likely cause of this difference? *1) Harder questions on Y's test 2) Better teaching by X 3) Smarter students in X's class	47
17. For better or worse, Professor X has decided to give 15% A's, 25% B's. 40% C's. 15% D's and 5% F's. Given this circumstance, which of the following should a student prefer for a final exam from Professor X? *1) An exam with an average score of 50%-60% 2) An exam with an average score of 70%-80% 3) An exam with an average score of 90%	38
18. In nearly every course, some things are relatively easy to learn and are mastered by nearly all students. To what degree should this material appear on tests? *1) It should not be included. The instructor should assume everyone knows it and use testing time to evaluate their performance on the more difficult material. 2) Such material should be included because it makes up a substantial percentage of what is to be learned. In its absence on a test, scores will be unfairly low.	6
19. Which of the following is correct about the large majority of tests given in college classes? 1) The scores reveal approximately what percentage of course material each student knows. *2) The scores rank the students approximately according to how much course material they know.	61
20. The average QCA at Harvard and at Virginia Tech are not very different, yet admission standards differ radically. How can this be accounted for? 1) Students learn about as much at Virginia Tech as at Harvard, though it's probably easier for most students at Harvard. *2) Standards for grading are higher at Harvard than at Virginia Tech.	69
21. Professor W's exams are usually taken directly from the text and lectures, while those of Professor Z usually involve problem solving or even require mastering new concepts presented in the exam itself. Which is more fair? 1) W's exam *2) Z's exam	13

Table 2B is continued on the next page.

Table 2B Continued

Questions 16-25 Continued		Percentage "Correct"[1]
Professor Q gave a final test of 20 questions to her 37 students in introductory geography. Like many college tests, the test covered a range of material. The resulting score distribution is shown in Figure 1 below. For each of the following conclusions, blacken answer space 1 if the conclusion is justified or space 2 if the conclusion is not justified.		
22. Most of Professor Q's students knew more than half of what was taught	*(2)	22
23. The substantial proportion of scores below 70% suggest that something was wrong educationally. (Perhaps the students were not well prepared or the instructor did not teach well.)	*(2)	39
24. Suppose Professor Q had decided in advance to give 15% A's, 25% B's, 40% C's, 15% D's and 5% F's regardless of how the scores turned out. In this case, the test would have been more fair if it had been harder, say with an average score of 55%.	*(1)	45
25. A student who scored 20 knows about twice as much geography in the area tested as a student who scored 10.	*(2)	74

Figure 1



[1]Asterisks indicate "correct" answers.

Table 3

Textbooks' Treatment of Test Difficulty

Textbook	Test Difficulty Topic													
	Type of decision influences desired difficulty level	Student ability affects difficulty level	Difficulty level determines spread and discrimination	Difficulty level defined	Easy item defined	Hard item defined	Disadvantages of easy or hard items	Recommend levels of difficulty for tests	Intercorrelations among items and difficulty levels	Procedures for estimating difficulty levels	Recommend ordering items by difficulty level	Item difficulties determine mean percentage-correct score	Interpretation of percentage-correct test scores	
Ahmann & Glock, 1971	+	-	+	+	+	+	+	+	+	+	+	-	-	
Anastasi, 1968	+	+	+	+	+	+	+	+	+	-	+	-	-	
Chase, 1974	+	-	+	+	+	+	+	+	?	+	-	-	-	
Cronbach, 1960	+	-	+	-	-	-	-	-	-	-	-	-	-	
Ebel, 1972	-	+	+	+	+	+	+	+	-	-	+	+	+	
Gerberich, Greene & Jorgenson, 1962	-	-	-	+	+	+	+	+	-	+	-	-	-	
Gronlund, 1976	+	+	+	+	+	+	+	+	-	+	+	+	-	
Gronlund, 1977	+	?	+	+	+	+	+	+	-	+	+	+	-	
Horrocks & Schoonover, 1968	-	+	-	+	+	+	-	-	-	+	+	-	-	

Note: Key to entries in Table 3:

+ = Topic discussed

- = Topic not discussed

? = Could not judge

Table 3 is continued on the next page.

Table 3 (cont.)

Textbooks' Treatment of Test Difficulty

Textbook	Test Difficulty Topic												
	Type of decision influences desired difficulty level	Student ability affects difficulty level	Difficulty level determines spread and discrimination	Difficulty level defined	Easy item defined	Hard item defined	Disadvantages of easy or hard items	Recommend levels of difficulty for tests	Intercorrelations among items and difficulty levels	Procedures for estimating difficulty levels	Recommend ordering items by difficulty level	Item difficulties determine mean percentage-correct score	Interpretation of percentage-correct test scores
Lien, 1976	+	-	+	-	+	+	+	+	-	+	+	-	-
Marshall & Hales, 1971	+	+	+	+	+	+	+	+	+	+	+	+	-
Martuza, 1977	+	+	+	+	+	+	+	+	-	+	+	-	-
Mehrens & Lehmann, 1973	+	+	+	+	+	+	+	+	+	+	+	-	-
Noll, 1965	+	-	+	?	?	?	+	+	-	-	-	+	-
Sax, 1974	+	+	+	+	+	+	+	+	+	-	-	+	-
Stanley & Hopkins, 1972	+	+	+	+	-	-	+	+	-	+	+	+	+
TenBrink, 1974	+	+	+	+	+	+	+	+	-	+	+	+	-
Thorndike & Hagen, 1961	-	-	-	-	+	-	-	-	-	+	+	-	-
Tuckman, 1975	-	+	-	+	-	-	+	+	-	+	+	-	-

Note: Key to entries in Table 3:
+ = Topic discussed
- = Topic not discussed
? = Could not judge

App. A

College Achievement Testing: A Self-Instructional Interlude

by Robert B. Frary and Robert W. Bowman, Jr.

A Parable as Prologue

Once upon a time, a great king had grown weary of the ineffectiveness of some of his knights to defend him adequately. The king decided to hold a test to see who among the knights should be admitted to the inner circle, who should be assigned to lesser stations, and who should be released from service. The king's privy council developed and administered the test. In good time they presented the scores to the king.

At first, the king was pleased. The knights had done well on the test; the average score was 80 per cent. When it came time to classify the knights, however, the king noted with dismay that the scores were so bunched together that it was hard to discriminate among them. It was particularly difficult to decide whom to admit to the inner circle. The king knew from his own experience, for example, that Sir Edelbert was in all ways superior to Sir Montrose, yet both knights were among 25 who had made perfect scores of 100 per cent.

Fortunately, with the king's greatness had come recognition of his own limits, and he consulted the realm's foremost philosopher concerning his predicament. The philosopher listened

to the king with interest, and then she spoke:

"Sire, you require a wider distribution of test scores, one in which fewer cases occur at each point. Then you can decide with greater confidence, for example, to admit Sir Edelbert to the inner circle and to assign Sir Montrose to the second circle. And you will be able to justify your decision to them."

The king thought for a moment and asked, "Yes, wise one, how can this be achieved?"

"Your majesty, look at this question on the test, about the use of the battle axe. Notice that all 250 knights correctly answered this question. All told, nearly half the questions on the test were answered correctly by nearly all the knights."

"Yes," said the king, "isn't it wonderful our knights are so knowledgable?"

"Perhaps," said the philosopher, "but what does the question about the battle axe tell you about the relative knowledge of your knights? Does it help to decide which knights to assign to which stations?"

"No, we guess not. Questions such as these give no information which helps to discriminate among our knights."

"So," the philosopher continued, "not only do these easy questions fail you in your decision, but they waste the resources of your treasury. Humbly I must remind you such waste is improper in these difficult times."

"Yes, you are right," said the king, "but tell us how these improvements may be achieved!"

Smiling, the philosopher answered, "Make the test harder,"

and she instructed the great king in her lore.

What the King Learned: A Modern Version

In what follows, the philosopher's presentation is updated and adapted (tongue-in-cheek) to a programmed instruction format. To follow along, simply read and answer each question and check your answers.

1. A teacher of a freshman English composition course was concerned about the large numbers of misspellings by students. To assess the situation further, he selected at random 100 words from a list of the 20,000 most commonly used words in written English and administered them orally as a spelling test.

   Suppose a student spells correctly 95% of the words on the test. What can you conclude about this student?

   a. The student can spell correctly about 19,000 of the 20,000 words on the list.

   b. The student will misspell about 5% of the words on a typical assignment.

ANSWER: The correct answer is a (assuming the instructor's pronunciation was adequate). Answer b is wrong because the range of words in the student's compositions probably does not coincide very well with

the list of 20,000. This student may well have less than a 5% misspelling rate due to avoidance of unfamiliar words. Go to question 2.

2. Another English instructor had the same concerns that motivated the test described above. However, this instructor selected 100 of the more commonplace words from a list entitled, "The 500 English Words Most Frequently Misspelled." Which instructor probably got more useful information?

   a. The first instructor

   b. The second instructor

ANSWER: The correct answer is b. It is true that the first instructor knows the approximate level of each student's spelling ability on an absolute scale. However, there were probably so few difficult words on the test that luck in the choice of these words might affect a student's score substantially. Suppose, for example, that the instructor decided to require additional spelling practice for those below a given standard. There is a good chance that a number of students would be misassigned due to the choice of the few difficult words being lucky or unlucky for them.

Another problem with the first instructor's approach is that it provides very little information concerning the kinds of words students are likely to misspell. Probably many words on the test were

spelled correctly by all or nearly all students, which provides little useful information. Go on to question 3.

3. A history instructor teaching freshmen for the first time observed apparent widespread ignorance in areas usually taught in high school. To check the extent of this condition he selected 100 germane questions from teachers' manuals for three widely used high school history texts and administered them to his students. Their average score was 55%. What can the instructor conclude?

   a. The students are indeed ill-prepared (on the average). One should expect average college bound high school seniors to be answering at least 85% to 90% correct while still in high school, in which case a drop to 55% seems excessive.

   b. The instructor knows little more than before the test with respect to the average knowledge level of the class. However, it is possible to identify the weaker students. Also, the content of specific questions and the proportion answering each of them correctly may give clues to areas of relative strength or weakness.

ANSWER: If you chose answer b you are correct! Answer a is incorrect. The difficulty of questions on most academic tests is dependent on a great many factors other than their basic content. In the case of

multiple-choice tests, the attractiveness of the distractors is a major determiner of question difficulty, for example,

What is the capital of Pennsylvania?

a. Harrisburg		a. Harrisburg
b. Pittsburgh	vs.	b. Reading
c. Philadelphia		c. Lancaster
d. Erie		d. Bethlehem

Go to question 4.

4. The history test just described was multiple-choice with four choices per question. The lowest scoring student in the class got 30% (30 out of 100 correct). What can be said of this student?

   a. A student guessing completely at random would get about one out of four questions right or 25%. Therefore, the score of 30% must represent something closely akin to total ignorance.

   b. The student in question probably knows less in the area tested than any other student in the class, but no statement can be made about the absolute level of knowledge.

ANSWER: You're right if you chose b. Answer a is wrong. It is true that a completely ignorant examinee will probably score near 25% or even as high as 30% if lucky. Further, there may even be completely ignorant examinees in the class, although this would seem far-fetched in most academic settings. Nevertheless, it is entirely possible that an examinee with a

reasonable amount of knowledge will score as low as 30% even in the absence of transitory personal problems that might depress performance. It is necessary only that the test contain sufficiently difficult questions or that the examinee be sufficiently misinformed. Indeed, if the examinee believes the answer to a multiple-choice question to be a wrong response, the best of luck will not yield a score point.

Go to question 5.

5. A course, entitled Introductory Psychology, is taught at two fully accredited universities. Inspection of the syllabi and course grade distributions reveals:

   Course content - virtually identical

   Text - same

   Grading standards - identical percentage ranges for grades based on averaging multiple choice test scores (93% - 100% = A, etc.)

   Grades actually awarded - Nearly equal proportions of A's, B's, C's, etc.

   However, one school is highly selective and is known for maintaining rigorous academic standards; the other has open admissions and is not known for high standards. What can be said about the similar achievement records for Introductory Psychology?

   a. The school with open admissions must be doing a better job of instruction, since, obviously, students of

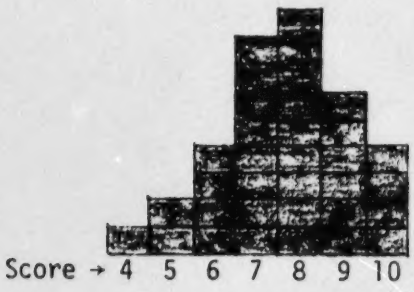
lesser ability are obtaining scores as high as academically talented students at the selective institution.

b. This situation almost surely reflects the presence of substantially easier test questions (on the average) at the open admissions school.

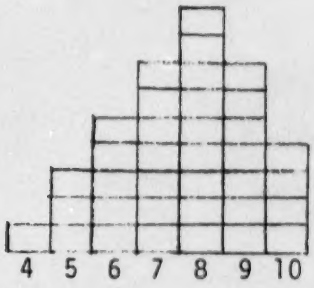
ANSWER: Answer b is correct, at least provisionally. The actual answer to this question would have to be determined empirically, perhaps by cross-administering the tests from the two schools. At the least, however, it is certainly possible and, indeed, plausible that easier questions account for the observed outcome. Choosing a suggests a recurring tendency to view all tests as similar to the spelling test in Question 1. Instead, the difficulty of a psychology test depends largely on the content of the questions selected and on the way they are stated.

Go to question 6.

6. Shown below are the score distributions for two, ten-item, Introductory Psychology quizzes on the same topic at the two schools described in 4. (Each box represents a student.)



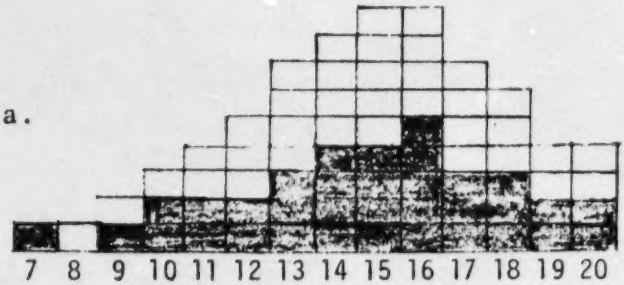
Open N=34



Selective N=36

Suppose that the two quizzes had been merged to make a single 20-item quiz and that this had been administered at both schools. What would the combined score distribution probably look like?

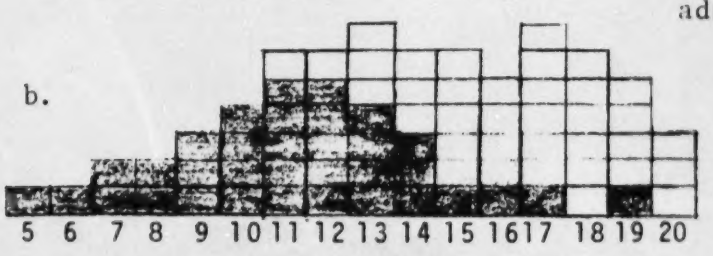
a.



Open and Selective Combined N=70

Shaded boxes represent students from the open-admissions school.

b.



ANSWER: Choice b is the correct answer. If you chose a, you should remember that the questions are probably much easier at the open-admissions school. Therefore, the students at the selective school will miss very few of these questions. Most of their scores will be 8 to 10 points higher on the combined quiz. In contrast, most of the open-admissions students will gain very few points from among the 10 added to their tests. Note, however, that in answer b above one open-admission student scored 95% on the combined quiz and one student at the selective school scored only 50% on the combined quiz.

Go to question 7.

7. Suppose that the test merging just described was actually carried out as an experiment and that, while somewhat concerned about the results, both schools decided to continue giving about the same proportions of A's, B's, etc., as before. Under this condition, which would be more fair to the students?

   a. Continue to use the same kinds of tests as before.

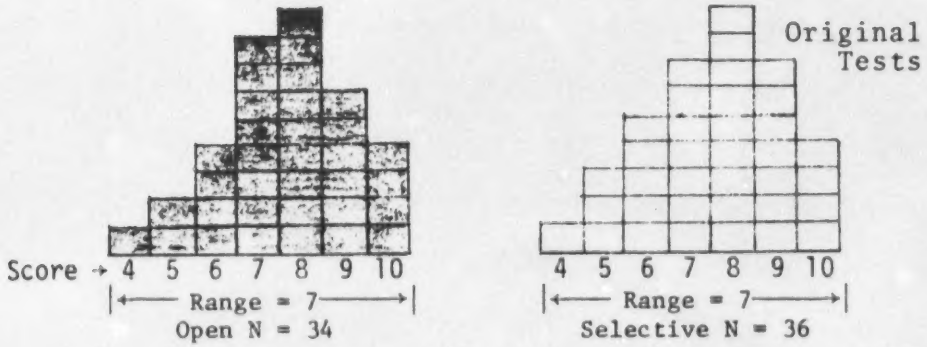
   b. Use tests with wider ranges of difficulty.

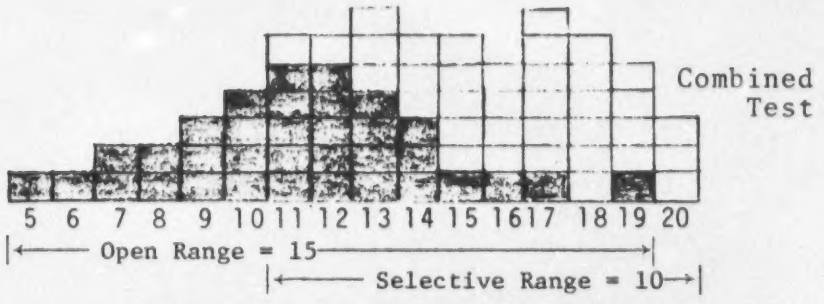
ANSWER: Choice b is correct, but this answer requires some qualification. The merged test would serve much better for the open-admission school in comparison with their original test. Suppose 15% (five students)

are to receive A's. In the original distribution four students received 10 points and six received 9. Probably the decision would be made to award A's only to the four 10's, thus, in a sense, shortchanging all six 9's. For the merged test, the top four students would again get A's, but only three students occur in the next lower category, thus reducing the number subject to an adverse decision. This example is trivial in itself but illustrates the advantage of a wider distribution of test scores. Fewer cases occur at any decision point thus lessening the effect of unavoidable erroneous classification decisions. The goal then of testing (in most academic situations) should be to spread the examinees out as much as possible over the available score range.

Please turn the page and continue reading.

Observe again the original and combined score distributions from the example above.





The distribution of the scores from the selective school on the merged test was spread out little more than before (a range of 10 points versus 7 previously). Compare this to the change in range for the open-admission scores (15 points versus 7). What was really needed for the selective school was a harder test. Probably some but not all of their top

scorers (9's and 10's) could have answered more difficult questions and thus more effectively established their right to receive high grades.

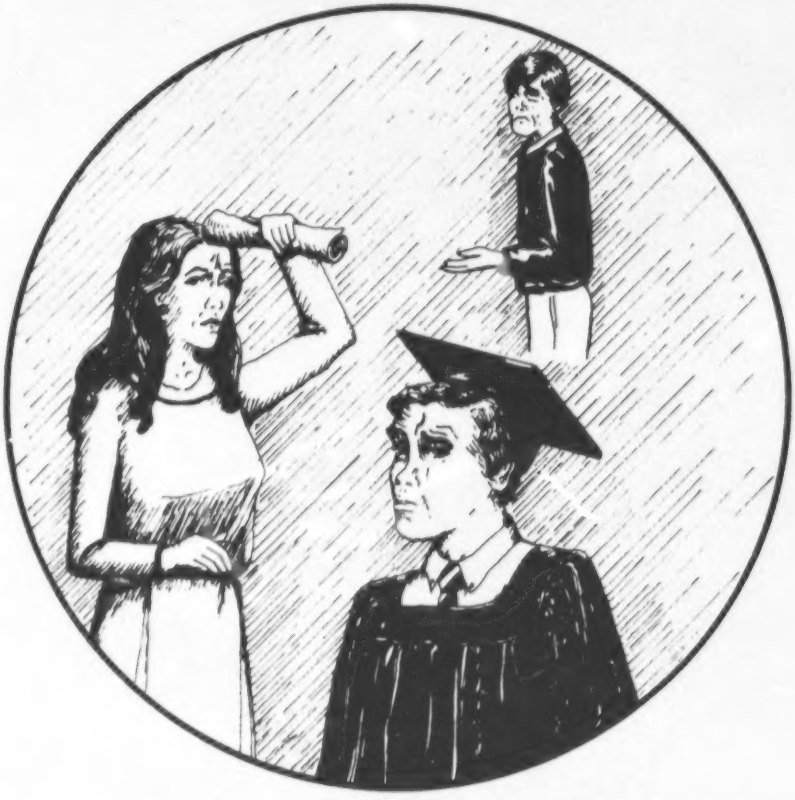
Note that the above argument, while framed in terms of evaluating high scorers, actually applies to the entire range of scores.

8. If academic tests are made hard enough to yield scores ranging from say 25% to 95%, that would result in average scores of about 60%. In other words, half the students will make less than 60%. How can you assign a C, for example, to a score of 45% or even 50%?
   a. You can't. This outcome clearly implies the student knows less than half of what was taught, and that's not what a C indicates.
   b. All that is required for this approach to testing and grade assignment is that those involved be informed. Students should know, for example, that tests need to be difficult in order to have good measurement accuracy. The old idea that 70% should be passing just doesn't apply outside of elementary school.

ANSWER: If you chose answer b, you are correct! Now you share the king's (and the philosopher's) wisdom. If you chose answer a, it's not too late: You may return to Question 1 and work through the program again.

App. B

The Trouble With Headaches



A Horrible Romance

Linda and Kevin met in the hall after Dr. Johnson's class. Kevin could tell that Linda was worried-- probably about the multiple-choice test they had just finished. Linda was comparing her answers with the answer key given out after class. Sympathetically, Kevin asked, "What's the problem, Linda?"

"I studied for this test and I never get less than B's, but I don't think I got over 60%," Linda said with dismay. "I did have a headache, but that couldn't have made me do this badly!"

"It'll be okay; Dr. Johnson said he'd give the same numbers of each letter grade regardless of the scores," soothed Kevin. "That's the way all the profs did it at Michigan."

"Well, this is Virginia Tech, so maybe you'd better explain some more."

What's happening here?

Think back over the tests you have taken in college. Do they remind you of a dog's breakfast-- lots of soft, easy questions, some tougher, chewy questions and a few indigestible lumps? But mostly soft.....?

After all, we grew up taking spelling tests where a score of 70% meant we knew about 70% of the words on a long list from which the test words were a representative sample. Unfortunately, this simple approach to testing breaks down even in elementary school. How could you get a list of all the questions that could be asked in a history course? And even if you could, selecting randomly from the list probably wouldn't make a satisfactory test. In fact, just as in making a dog's breakfast, most college teachers don't really choose how to construct tests. They just sit down and write the best questions they can to cover what they have taught. Eventually, with experience, they learn to make their tests about as difficult as they intend. For many, that means that a student who scores 70% will receive a D.

Since 70% is absolutely arbitrary, we should question whether it is desirable. In the discussion to follow, it is argued that a 70% passing level is <u>much too high</u> for a large majority of college tests. There are exceptions of course; this claim does not apply to essay tests or tests designed to measure a specific skill. What it does concern are tests covering fairly diverse topics on which scores are determined by adding up points for each right answer. Almost any multiple-choice midterm or final examination would be in this category as would most short answer and problem solving tests.

Since these tests may be as hard or as easy as the teacher happens to make them, it is clear that the percentage of right answers does not estimate some level of achievement directly as do elementary spelling tests. What the scores do provide is a ranking of students in terms of their achievement in the area represented by the questions. Consider a typical midterm with 50 questions. If nearly all the scores are in the range of 35 to 50 points (or 70% to 100%), there are only 15 points over which to rank the students for the purpose of determining at least four letter grades, D through A. At each cut point, there will be a number of students who narrowly gain the higher grade or barely miss it. A greater spread of scores may be obtained if the test is made much harder. Those who made 35 to 50 on the easy test may have scores ranging from 20 to the high forties on a much harder test. Then there will be many fewer students with scores adjacent to each cut point for determining letter grades. This spreading of scores minimizes the effect of minor scoring errors or variation in grading leniency. Assuming that the same proportion of A's, B's, C's, etc., are given as before, the result is a much more accurate assessment of achievement.

Think of all this in terms of Linda's headache. Suppose she missed two questions she would not have had she felt better. With the scores spread out, it is less likely that the two-point loss carried her over a grade boundary. Suppose Dr. Johnson had a headache too and made a lot of two-point errors in scoring the papers! Again, his mistakes are much less likely to affect grades if the cut points are far apart.

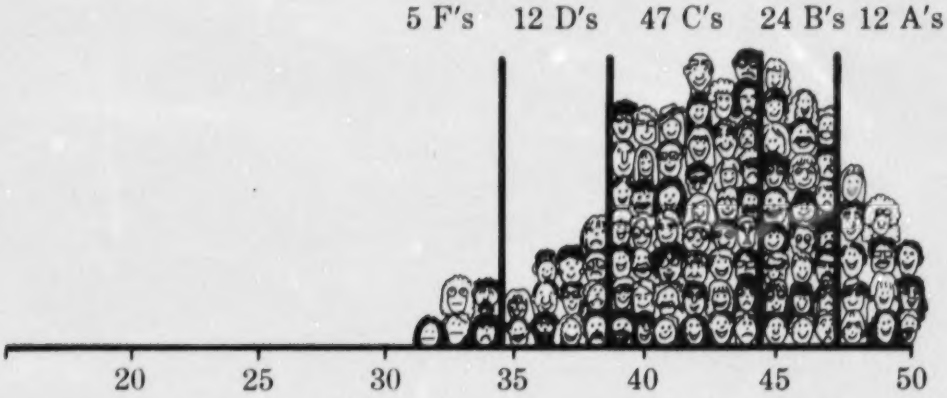
At this point, the clever reader will see that a 50-question test with all the scores between 35 and 50 must have about 35 really soft, easy questions. Does that mean that a teacher should throw out all the questions that nearly everyone can answer? No, probably not all, for the same reason that a musician plays a few scales before performing or that a broad jumper takes some practice hops before she takes the big jump. But, once the warmup is over and the real event (that means the test) has begun, it is more efficient, humane and equitable to ask those questions most likely to yield the widest spread of scores over the available range.

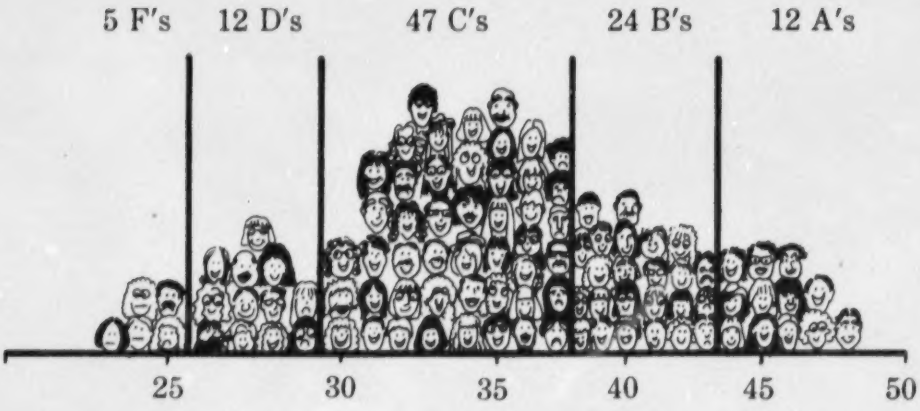
The teacher who introduces harder tests will have a little adjusting to do with respect to grade assignment. Obviously, if the average score is about 55%, the lowest passing score will have to be somewhat below 50%. This outcome will bother some people, though it shouldn't. After all, the easy questions just weren't asked, and it is reasonable to assume that most students would have answered them correctly.

Then nobody has to worry about getting a headache!

(See illustration, over.)

Scores of 100 students on hard and easy versions of a 50-question test. Count the sad faces.





Copyright © 1983 by Robert W. Bowman, Jr., and Robert B. Frary, Learning Resources Center, Virginia Polytechnic Institute & State University, Blacksburg, VA 24061. Artwork by Cindy Newberry.